DOCUMENT RESUME

ED 367 670 TM 021 107

AUTHOR Beasley, T. Mark
TITLE Education as a Fixed Effect Fallacy.
PUB DATE Feb 94
NOTE 22p.; Paper presented at the Annual Meeting of the Eastern Educational Research Association (17th, February 12, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Analysis of Variance; Behavioral Science Research; *Educational Research; Equations (Mathematics); Error of Measurement; Models; Research Problems; *Robustness (Statistics); Statistical Distributions
IDENTIFIERS *Fallacies; *Fixed Effects; F Test; Random Effects

ABSTRACT

In educational research, nonessential factors are commonly ignored and when accounted for, they are often treated statistically as fixed effects. Yet many researchers in these situations generalize their findings beyond the specific levels selected; however, the analyses may require treating the factor as a random effect. Such inappropriate analyses create the potential for an "education-as-fixed-effect fallacy." The present paper takes Clark's (1973) comments on the "language as fixed-effect fallacy" and applies them to educational research. The distinction between fixed and random models and the use of the quasi-F statistic are reviewed. Examples of commonly used mixed models with derivation of expected mean squares are provided and general guidelines for defining factors as fixed or random are suggested. (Contains 30 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 367 670

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN **OTHER THAN PAPER COPY** HAS BEEN GRANTED BY

T. MARK BEASLEY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Education as a Fixed Effect Fallacy

T. Mark Beasley

St. John's University

Paper presented at the meeting of the
Eastern Educational Research Association
February 12, 1994

Direct all correspondence to:
T. Mark Beasley
DAIL, School of Education
St. John's University
Jamaica, NY 11439
beasleyt@sjuvm

TM021107

Abstract

In educational research, nonessential factors are commonly ignored and when accounted for, they are often treated statistically as fixed effects. Yet many researchers in these situations generalize their findings beyond the specific levels selected; however, the analyses may require treating the factor as a random effect. Such inappropriate analyses create the potential for an *education-as-fixed-effect fallacy*. The present paper takes Clark's (1973) comments on the *language as-fixed-effect fallacy* and applies them to educational research. The distinction between fixed and random models and the use of the quasi-*F* statistic are reviewed. Examples of commonly used mixed models with derivation of expected mean squares are provided and general guidelines for defining factors as fixed or random are suggested.

Typically, behavioral researchers have simply ignored potential violations of assumptions when using the analysis of variance (ANOVA) even though most have been warned and are aware of the possible consequences. One reason for this is the common view that the ANOVA is robust to such violations. The Central Limit Theorem (CLT) states that if two populations have finite variances σ^2_1 and σ^2_2 and means μ_1 and μ_2, then the sampling distribution of means for their respective samples of n_1 and n_2 independent observations approach normal distributions with variances of σ^2/n_1 and σ^2/n_2 and means of μ_1 and μ_2 as the sample sizes increase. Moreover, as these sample sizes increase the sampling distribution of the difference between sample means approaches a normal distribution, regardless of the shape of the two population distributions (Wampold & Drew, 1990). This property of the CLT has major implications for the use of the ANOVA. Specifically, the fixed-effects ANOVA is not only a comparison of sample means, but it also involves inferences about populations means. Thus, the fixed-effect ANOVA is a test that compares the variance of the distribution of differences in sample means to an aggregate of squared standard errors, σ^2/n_i which estimates the population error variance. Therefore, the CLT implies that if sample sizes are sufficiently large the fixed-effects ANOVA will be robust to violations of the normality assumption.

Empirical evidence of the robustness of the ANOVA to violations of the normality assumption has been provided (Boneau, 1960; Box, 1954; Glass, Peckham, & Sanders, 1972; Lindquist, 1956). However, this long-standing view has recently been called into some question, by virtue of more sophisticated Monte Carlo studies conducted with more complicated designs, and more simulation samples (e.g., Rogan & Keselman, 1977; Tomarkin & Serlin, 1986; Wilcox, Charlin, & Thompson, 1986). Furthermore, the CLT and most of the research into the properties of the ANOVA has been concerned with inferences about means, that is fixed-effects, rather than inferences about variances, random-effects, which theoretically are not as robust to normality violations (Scheffe', 1959).

Unfortunately, many behavioral researchers and other consumers of statistical knowledge (i.e., graduate students) are not fully aware of the distinction between fixed and random effects models. In many research situations, variables that are implicitly generalized as random effects are commonly treated statistically as fixed effects and often even ignored. In semantic memory research, Clark (1973) commented on the "language-as-fixed-effect fallacy." He observed that words used as experimental stimuli were often analyzed as fixed effects when the researcher eventually generalized the findings (i.e., significant fixed factor treatment effects)

to the whole language (population of words). This type of generalization necessitates treating the stimuli factor as a random effect. Those committing this error, however, should only make inferences to the target population of subjects under the conditions of using the *same* treatment and the *identical* set of stimulus words. Fixed-effect fallacies have also been noted in research in which adjectives in personality prototype research (Richter & Seay, 1987), age levels in lifespan research (Bell, 1992), and messages in human communication research (Dawes, 1983) were used as control factors.

This design consideration may also pertain to pedagogical inquiry, that is an *education-as-fixed-effect fallacy*. In education, researchers typically devise new instructional techniques or methods and administer them in a standard experimental design with students (or classrooms, schools, etc.) as the only random factor(s). The findings are then generalized to the population of students but also across other factors (i.e., instructional units, curricular materials, and intra-curricular content) which were either not accounted for or were treated as fixed-effects. Yet, generalization should only be made to other students (random-effects) using the specific instructional and environmental stimuli used in the experiment. For example, a new instructional technique may be administered to two experimental classes that happen to be learning two separate units (e.g., fractions and multiplication), while two control groups happen to be studying decimals and division, respectively. But, one would not conclude that the effects of this technique generalize to students learning algebra, geometry, or mathematics in general. Yet in the same manner, instructional units and other *nonessential* factors are often ignored or treated as fixed effects. Such inappropriate analyses not only bias the estimate of the treatment effect, but also limit the generality of the findings. The underlying issue is that unless the researcher is doing a summative evaluation, generalization to other contexts is of interest. Variables such as instructional content which are not typically manipulated nor of theoretical interest, but have some effect on the dependent variable, should be treated statistically as random so that treatment effects can be properly decomposed and generalized to other contexts. Furthermore, the view that the administration of instruction works in an all-or-none fashion may also constitute an education-as-fixed-effect fallacy. For example, a new instructional technique may be used as a supplement for one hour biweekly by the experimental group and not at all by the control group. The treatment in this case, however, represents only one of the number of possible treatments that could be administered and of educational interest. Yet it does not technically constitute a

random-effect because treatments were not randomly sampled from the population of possible treatments. These types of situations make it difficult for the researcher to decide whether a factor should be treated as fixed or random. Moreover, in more complex mixed designs[1], the analysis of fixed and random effects requires the use of a distributional approximate, the quasi-*F* statistic. Thus, the two major debates underlying the acknowledgment of random effects in educational research concern the definition of "random effects" and the statistical properties of the quasi-*F* approximate statistic that is necessary to analyze random effects in many mixed designs. The current paper reviews the difference between random and fixed effects, reviews the research completed on the robustness and power of the quasi-*F* statistic, and investigates the issue of defining "random effects" in educational research from the perspectives of Clark and his opponents.

Distinction Between Fixed and Random Effects Models

In the ANOVA, a factor is considered *fixed* if the levels of the independent variable are selected on an arbitrary and systematic basis (Keppel, 1982). Therefore, the selected levels are assumed to represent the entire population of theoretically interesting treatment conditions, and thus, statistical generalizations are limited to the *J* levels of the fixed factor used in that particular analysis. The linear model for the one-way, fixed effect ANOVA is

$$Y_{ij} = \mu_{*} + \alpha_j + e_{ij}, \tag{1}$$

where *j* refers to the *J* levels of the fixed effect (α) and *i* refers to the *n* subjects in each of the *j* levels. The mean fixed treatment effect,

$$\sum_{j=1}^{J} \alpha_j = \sum_{j=1}^{J} (\mu_j - \mu_{*}) = 0, \tag{2}$$

is equal to zero because the *J* levels of α represent the entire population. Therefore, the estimate of the population grand mean is based in the observations and is calculated as the average of the *J* group means. It should be noted, however, that the sum of the squared treatment effects equal zero,

$$\sum_{j=1}^{J} \alpha_j^2 = 0, \tag{3}$$

only if the null hypothesis is true. When the null hypothesis is not true,

$$\sum_{j=1}^{J} \alpha_j^2 > 0, \tag{4}$$

and a non-centrality parameter (NCP) is estimated with the *F* test. That is, an estimate of the fixed factor treatment effect, based on the expected value of the mean squares between groups (MSF),

$$E(MSF) = \sigma_e^2 + \frac{\sum_{j=1}^{J} n_j \alpha_j^2}{J - 1} = \sigma_e^2 + n\theta_\alpha^2 , \quad (5)$$

is divided by an estimated error term based on the expected mean square error (MSE),

$$E(MSE) = \sigma^2_e . \quad (6)$$

If the null hypothesis is true, then the *F* ratio,

$$F = \frac{MSF}{MSE} , \quad (7)$$

has a distribution with $(J - 1)$ and $(N - J)$ degrees of freedom. If the null hypothesis is false, not only is the treatment effect greater than zero, but the expected mean square error is greater than the population error variance

$$E(MSE) > \sigma^2_e . \quad (8)$$

The ratio of these biased estimates forms an *F* statistic with $(J - 1)$ and $(N - J)$ degrees of freedom which is compared to the *F* distribution to estimate the probability of such a result occurring by chance.

In contrast, a factor qualifies as *random* if the levels of the independent variable are selected (i.e., sampled) randomly or unsystematically from a larger pool of possible treatment conditions (Myers, 1979). The selected levels are assumed to represent a random sample of the population of treatment conditions and the statistical generalizations extend beyond the levels of the independent variable included in the analysis. Computationally, the one-way, random-effects ANOVA is identical to the fixed effects model (1),

$$Y_{ij} = \mu_\bullet + A_j + e_{ij} ; \quad (9)$$

however, the uppercase Roman letter (A) instead of the Greek symbol (α) is used here and subsequently to distinguish between random and fixed factors. The only restriction is that if *J* levels of the random factor (A) are sampled, then there must be at least $(J + 1)$ treatments, else a fixed-effect is being analyzed. In the random-effects ANOVA, the mean of the set of all treatments is zero because

$$\sum_{j=1}^{J} A_j = \sum_{j=1}^{J} (\mu_j - \mu_\bullet) = 0 ; \quad (10)$$

however, the mean of the effects for those treatments selected and actually involved in the analysis typically does not equal zero,

$$\sum_{j=1}^{J} A_j = \sum_{j=1}^{J} (\bar{a}_j - \mu_\bullet) \neq 0 . \quad (11)$$

Importantly, the variance of A is zero if and only if the effects of all possible treatments are zero. Hence, the null hypothesis for the random effects ANOVA is that the variance of A equals zero:

$$\text{Ho: } \sigma^2_A = \sum_{j=1}^{J} (\mu_j - \mu_\bullet)^2 = 0 \ . \qquad (12)$$

The *F* ratio under conditions of a true null hypothesis is

$$F = \frac{\text{MSR}}{\text{MSE}} , \qquad (13)$$

and has an *F* distribution with $(J - 1)$ and $(N - J)$ degrees of freedom; however, the expected value of MSR differs from MSF in that the treatment is estimated as an expected variance

$$\text{E(MSR)} = \sigma^2_e + n\sigma^2_A. \qquad (14)$$

Thus, although in the one-way ANOVA random and fixed effects are estimated in identical manners, the hypothesis being tested is different. Specifically, fixed effects involve inferences about mean differences (3); whereas, random effects involve inferences about variances (14). Furthermore given the condition stated in (11), the CLT applies to the *sampling of levels* from the population of treatment conditions *not to the sampling of subjects* under those treatment conditions. Therefore, increasing the number subjects in a random effects design does not guarantee that the estimated grand mean will approach the population grand mean or that the distribution of sample means approaches a normal shape. Thus, the random effects ANOVA is not theoretically expected to be as robust to normality violations as the fixed effect design. Specifically, nonnormality in the form of kurtosis deviating from zero affects the Type I error rate and power of the *F* ratio for random effects. As kurtosis exceeds one and the distribution becomes leptokurtic (i.e., spiked) the Type I error rate exceeds the nominal alpha. In an opposite manner, when kurtosis is less than zero and the distribution becomes platykurtic (i.e., flat) the nominal alpha becomes a "conservative" estimate of the actual Type I error rate. Thus, although violating the normality assumption may have slight effects for fixed factors in the ANOVA (i.e., Glass et al., 1972), random effects are severely affected by distributional deviations from normality. The reason for this is that fixed effects involve inferences based on the deviation of an estimate (group mean) from an internally estimated parameter (grand mean) which has a basis in the observations (2). Furthermore, the CLT makes provisions for the distribution of mean differences regardless of distributional shape. Inferences about variances, however, involve the deviations of estimates and their ratio to the population parameter (σ^2) which has no necessary basis in the observations (11). Therefore, the CLT makes no distributional provisions. Thus, a

biasing parameter such as kurtosis which affects the shape and variance of a sampled distribution also affects inferences about variances (Scheffe', 1959).

Mixed Designs with Crossed Random Factors

The distinctions between random and fixed factors have particular importance in mixed designs which have at least one fixed factor and at least one random factor. One of the most common mixed designs in educational research involves the repeated measures ANOVA (i.e., pre/post design) with subjects, a random factor (P), crossed with a fixed factor (α), say time of measurement[2]. The repeated measures ANOVA has the following linear model:

$$Y_{ij} = \mu_* + \alpha_j + P_i + \alpha P_{ij} + e_{ij}, \tag{15}$$

where j refers to the J levels of the fixed factor (α) and i refers to the n subjects (P). The expected mean square for the random factor, subjects (P), is

$$E(MSR) = \sigma^2_e + n\sigma^2_P; \tag{16}$$

however, the variance of the αP interaction cannot be separated from random error variance because all degrees of freedom are accounted for by the first four terms in (15). Thus, (16) can be expressed as

$$E(MSR) = \sigma^2_e + \sigma^2_{\alpha P} + n\sigma^2_P. \tag{17}$$

The expected mean square of the fixed effect also includes the time by subject interaction

$$E/MSF) = \sigma^2_e + \sigma^2_{\alpha P} + \frac{\sum_{j=1}^{J} n_j \alpha_j^2}{J - 1} = \sigma^2_e + \sigma^2_{\alpha P} + n\theta^2_\alpha . \tag{18}$$

Thus, the αP interaction is used as the error term for both the fixed (18) and random effects (17) and has the following expected mean square,

$$E(MSE) = \sigma^2_e + \sigma^2_{\alpha P}. \tag{19}$$

To extend this further, a mixed design similar to (15) in which the fixed factor (α; e.g., treatment vs. control) is crossed with a random factor (B; e.g., presentation order of instructional units), and subjects are nested in the αB interaction cells is shown

$$Y_{ijk} = \mu_* + \alpha_j + B_k + \alpha B_{jk} + e_{ijk}, \tag{20}$$

where, j refers to the J levels of the fixed factor (α), k refers to the K levels of the random factor (B) and i refers to the n subjects in each of the αB cells. The expected mean square for the random effect of presentation order (B) in (20) is

$$E(MSR) = \sigma^2_e + Jn\sigma^2_B; \tag{21}$$

however, because levels are randomly sampled from the population, the mean and variance of the random effect are biased and are not typically equal to zero (11). This random factor bias is crossed with the fixed factor, and thus, the interaction effect

(αB) may create differences in the treatment effects of α which would not occur if all levels of B were analyzed (Wampold & Drew, 1990). Thus, contrary to what one might expect intuitively, the fixed-effect rather than the random factor is contaminated by the interaction effect. Thus, the expected mean square of the fixed effect contains the interaction (αB),

$$E(MSF) = \sigma^2_e + \sigma^2_{(\alpha \times B)} + \frac{K \sum_{j=1}^{J} n_j \alpha_j^2}{J - 1} = \sigma^2_e + n\sigma^2_{(\alpha \times B)} + Kn\theta^2_\alpha , \quad (22)$$

and thus, uses the interaction mean square as an error term,

$$E(MS_{\alpha \times B}) = \sigma^2_e + \frac{\sum_{j=1}^{J} \sum_{k=1}^{K} n_{ij} (\alpha B_{jk} - \alpha_j - B_k + \mu_{\bullet\bullet})^2}{(J - 1)(K - 1)} = \sigma^2_e + n\sigma^2_{(\alpha \times B)} . \quad (23)$$

Since the random effect is not contaminated by the interaction, the mean square error, which has the following expected value,

$$E(MSE) = \sigma^2_e, \quad (24)$$

is used as the error term for the random effect (21) and the interaction (23). Notice that for except for the use of the symbol B instead of β, the model in (20) is identical to a two-way factorial ANOVA with both factors fixed in which case, MSE (24) is used as the error term for all effects. Thus, if the random effect (presentation order) were ignored or treated statistically as a fixed effects, then no bias in the estimation of the random or interaction effects (if analyzed) would occur. The effect of the fixed factor (α), however, would be overestimated due to the additional sources of error included in (22), which in turn would lead to more Type I errors, improper generalizations, and less robustness to normality violations.

In general, the effects of the inappropriate analysis of crossed random factors in mixed designs can be summarized as follows:

1). If the random factor (B) is treated as a fixed factor, then the fixed factor (α) will be overestimated because of the presence of the interaction (αB) in the expected mean square of α (22).

2). If the random effect is ignored, the interaction still contaminates the fixed factor, while the variance of the random effect is likely to be included in the error term. Thus, depending on the ratio of $\sigma^2_{(\alpha B)}$ to σ^2_B , (23) to (21), the fixed effect can be over- or underestimated.

Furthermore, even when random effects are analyzed appropriately, the effects of violating normality assumptions extend to mixed designs because the poorly estimated random-effect is crossed with the fixed factor, and thus, fixed effect estimates can be biased. This is because the interaction effect is contained in the

fixed effect (22) and although it is not typically affected by inappropriate analysis of these designs (Kirk, 1982), the random effect (21), which is not contaminated by the interaction (Wampold & Drew, 1990), is removed in the calculation of the interaction (23). The removal of a biased random effect may in turn bias the interaction and thus the fixed-effect.

Mixed Designs with Nested Random Factors

In some mixed designs, the random factor is nested in the fixed factor and thus contained in the expected mean square of the fixed effect[3]. In this situation, a random by fixed interaction cannot be calculated and the fixed and random effects cannot be directly separated. For example, a researcher investigating the fixed-effects (α) of prose presented in different formats (e.g., oral vs. written) on children's comprehension has the subjects (P) respond to each format in order to enhance the sensitivity of the design. Because of this within-subject factor and potential practice effects the passages (B) for each fixed factor group must be different. The subjects, however, respond to all passages in both prose formats. In this situation, the linear model is as follows:

$$Y_{ijk} = \mu_{\bullet} + \alpha_j + B_{k(j)} + P_i + \alpha P_{ij} + PB_{ik(j)} + e_{ij}, \quad (25)$$

where, j refers to the J levels of the fixed factor (α), k refers to the K levels of the random factor B nested in α, and i refers to the n subjects (P). The random factor, subjects (P) is crossed with both the passage random factor (B) and the prose format fixed factor (i.e., each subject responds to every passage in both conditions). Thus, the passage random factor (B) and the interaction of the subject random factor (P) with treatment (αP) and passages (PB) is contained in expected mean square of the fixed-effect:

$$E(MS\alpha) = \sigma^2_e + \sigma^2_{PB} + n\sigma^2_B + K\sigma^2_{\alpha P} + Kn\,\theta^2_{\alpha}. \quad (26)$$

Both of the random-effects and the subject by treatment interaction contain the random factors interaction in their expected mean squares,

$$E(MSB) = \sigma^2_e + \sigma^2_{PB} + n\sigma^2_B \quad (27)$$

$$E(MSP) = \sigma^2_e + \sigma^2_{PB} + K\sigma^2_P. \quad (28)$$

$$E(MS_{\alpha \times P}) = \sigma^2_e + \sigma^2_{PB} + K\sigma^2_{\alpha P}. \quad (29)$$

Thus, the interaction of these random factors is used as the error term for the F ratio to test the random effects,

$$E(MS_{P \times B}) = \sigma^2_e + \sigma^2_{PB}. \quad (30)$$

To test the fixed factor, however, no direct matches among the random effect mean squares exist such that the fixed-effect is isolated. Thus, the only way to estimate the

fixed effect is to construct a quasi-*F* statistic. Although two such formulations are possible, the following is more commonly used:

$$F' = \frac{MS_{\alpha} + MS_{P \times B}}{MS_{\alpha \times P} + MS_{B}}. \quad (31)$$

This and other "additive" quasi-*F*'s are preferred because they avoid the possibility of negative values. This ratio will be approximately 1.0 when the null hypothesis is true and will exceed one if the null hypothesis is false (Winer, 1971). This version of the quasi-*F* is distributed approximately as *F* if the degrees of freedom (*df*) are adjusted. The general rule is for the additive quasi-*F* is that for the numerator *df's*, the numerator of the *F* ratio is squared and then divided by the sum of their separate components squared divided by their original *df's*,

$$df_{num} = \frac{(MS_{\alpha} + MS_{P \times B})^2}{MS_{\alpha}^2/df_{\alpha} + MS_{P \times B}^2/df_{P \times B}}. \quad (32)$$

For the denominator, the inverse process holds, that is the denominator of the *F* ratio is squared and then divided by the sum of their separate components squared divided by their original *df's*,

$$df_{dem} = \frac{(MS_{\alpha \times P} + MS_{B})^2}{MS_{\alpha \times P}^2/df_{\alpha \times P} + MS_{B}^2/df_{B}}. \quad (33)$$

Thus, if the random effect is treated as fixed, then the quasi-*F* is not necessary; however, the fixed factor (α) is confounded with (B) and less conventional analyses are suggested (see Kirk, 1982). If the random factor is ignored in this situation, an overestimate of the fixed effect will occur because of the presence of the random effect variance, σ^2_B, in the expected mean square of the fixed effect (26).

Use of the Quasi-F statistic

Because of the quest for generality beyond the specific details of a particular experiment, Clark (1973) proposed that analyzing subjects and nonessential stimuli as random effects should be more common and recommended ANOVA models which are logically executable but have the disadvantage that many hypotheses cannot be tested with the standard *F* ratio. Mixed designs with two or more random factors often necessitate the use of a quasi-*F* ratio which will isolate the effect of interest and permit a test based on the standard linear model. Unfortunately, the exact distribution of the quasi-*F* is unknown. Under the conditions of a true null hypothesis, however, the *df*'s of the quasi-*F* can be adjusted (i.e., 32 and 33) so that it approximates the conventional *F* distribution (Winer, 1971). Although Clark's arguments had a tremendous effect on statistical analysis and design of language research, it has been virtually ignored in other areas of behavioral science (Richter

& Seay, 1987). One reason for the relative obscurity of the quasi-*F* test is a lack of information dissemination. Random effects models are not typically emphasized in graduate training much less the use of the quasi-*F*. Furthermore, random effect and mixed models tend to be more complex in their design, execution, and analysis (Wike & Church, 1976). Moreover, users of statistics from all areas have a general distrust of approximate statistics, no matter how approximate their measures may be. Approximate statistics, such as the quasi-*F* require unconventional adjustments and more importantly have relatively unknown statistical properties. However for the fixed effect (26) used in the previous design (25), the quasi-*F* statistic (31) has been shown to be robust to both the normality and variance homogeneity assumptions (Santa, Miller, & Shaw, 1979). Quasi-F ratios have also been shown robustness to violations of the sphericity assumption when within-subjects factors are analyzed (Maxwell & Bray, 1986). These estimates, however, tend to be conservative when the nested random factor (27) and the treatment by subjects interaction (29) variances are small (Davenport & Webster, 1973; Forster & Dickinson, 1976). A rather complex procedure to improve the Type I error rate and power of the quasi-*F* is available (Myers & Howe, 1971); however, the more logically sound approach is to purposely design experiments with larger numbers of stimuli in each condition in order to have sufficiently large variation for the random factor (Richter & Seay, 1987). Yet, some researchers will continue to be concerned about being penalized with the loss of power for performing the appropriate analysis. One must realize however, that the use of the quasi-*F* should be expected to be less powerful because extraneous variance in the treatment mean square is being controlled (31). Moreover, Santa et al. (1979, p. 45) suggest that in such cases, it is preferable to "break the stranglehold of $\alpha = .05$," than to persist in the misuse of *F* ratios in which *p*-values and alpha levels are meaningless. Thus, when the need for generality is greater than the desire for control and significance, the quasi-*F* is an appropriate alternative.

Examples of Commonly Used Mixed Designs that require the quasi-F

Nested Random Factor. In a variant of the previous design (25) a researcher comparing auditory and tactual teaching methods and their fixed effects (α) on children's' acquisition of number concepts may randomly assign subjects (P) to each condition because responding in both conditions may obscure the treatment effects. The test stimuli (B) for each fixed factor group are different (e.g., slides for the

auditory condition and objects for the tactual), although the subjects respond to all stimuli of a given type. In this situation, the linear model is as follows:

$$Y_{ijk} = \mu_{\bullet} + \alpha_j + B_{k(j)} + P_{i(j)} + PB_{ik(j)} + e_{ij}. \quad (34)$$

where, j refers to the J levels of the fixed factor (α), k refers to the K levels of the random factor B nested in α, and i refers to the n subjects (P). The random factors, subjects (P) and test stimuli (B) are crossed (i.e., each subject within a treatment condition responds to every test stimuli of that particular type). Thus both random factors and their interaction are nested in the expected mean square of the fixed effect,

$$E(MS\alpha) = \sigma^2_e + \sigma^2_{PB} + n\sigma^2_B + K\sigma^2_P + Kn\theta^2_A. \quad (35)$$

Both random effects contain their interaction in their expected mean squares,

$$E(MSB) = \sigma^2_e + \sigma^2_{PB} + n\sigma^2_B \quad (36)$$

$$E(MSP) = \sigma^2_e + \sigma^2_{PB} + K\sigma^2_P. \quad (37)$$

Thus, the interaction of these random factors is used as the error term for the F ratio to test the random effects,

$$E(MS_{P \times B}) = \sigma^2_e + \sigma^2_{PB}. \quad (38)$$

Again, the only way to estimate the fixed effect is to construct a quasi-F statistic,

$$F' = \frac{MS_{\alpha} + MS_{P \times B}}{MS_P + MS_B}. \quad (39)$$

The logic for adjusting the df's, also remains the same:

$$df_{num} = \frac{(MS_{\alpha} + MS_{P \times B})^2}{MS^2_{\alpha}/df_{\alpha} + MS^2_{P \times B}/df_{P \times B}}. \quad (40)$$

$$df_{dem} = \frac{(MS_P + MS_B)^2}{MS^2_P/df_P + MS^2_B/df_B}. \quad (41)$$

Crossed Random Factor. In the previous design, internal validity concerns prevented crossing the subjects with the fixed factor; however, in many cases another random factor may be crossed the fixed factor and the subject in order to increase design sensitivity. For example, subjects may be assigned to attended classes in one of two instructional methods: Team Accelerated Individualization (TAI) and Traditional Lecture Methods (TLM). The instructional units which serve as test units may be selected from a large pool of possible units. The same instructional units serve as tests for both conditions and the subjects in each group are tested on all units. Thus instructional methods, a fixed factor (α), is crossed with the random effect of instructional test units (B). Subjects (P) are nested within the fixed factor,

but crossed with the random unit factor. In this situation, the linear model is as follows:

$$Y_{ijk} = \mu_{\bullet} + \alpha_j + B_k + P_{i(j)} + \alpha B_{jk} + PB_{ik(j)} + e_{ij}. \quad (42)$$

where, j refers to the J levels of the fixed factor (α), k refers to the K levels of the random factor B crossed with α, and i refers to the n subjects (P). The expected mean square for the fixed effect contains the subject (P) effect and the interaction of units with treatment (αB) and subject (PB).

$$E(MS\alpha) = \sigma^2_e + \sigma^2_{PB} + n\,\sigma^2_{\alpha B} + K\,\sigma^2_P + Kn\,\theta^2_{\alpha} \quad (43)$$

The expected mean squares of the subject random effect and the instructional unit random effect (B) and the unit interaction with the fixed effect (αB):

$$E(MSP) = \sigma^2_e + \sigma^2_{PB} + K\sigma^2_P; \quad (44)$$

$$E(MSB) = \sigma^2_e + \sigma^2_{PB} + Jn\,\sigma^2_B; \quad (45)$$

$$E(MS\alpha \times B) = \sigma^2_e + \sigma^2_{PB} + n\,\sigma^2_{\alpha B}; \quad (46)$$

include the subject by instructional unit interaction which is used as the error term for testing their effects,

$$E(MS_{P \times B}) = \sigma^2_e + \sigma^2_{PB}. \quad (47)$$

To test the difference between instructional methods (α), the following quasi-F is constructed to isolate its fixed effect:

$$F' = \frac{MS_{\alpha} + MS_{P \times B}}{MS_P + MS_{\alpha \times B}} \quad (48)$$

The degrees of freedom are adjusted as follows:

$$df_{num} = \frac{(MS_{\alpha} + MS_{P \times B})^2}{MS^2_{\alpha}/df_{\alpha} + MS^2_{P \times B}/df_{P \times B}} \quad (49)$$

$$df_{dem} = \frac{(MS_{\alpha \times B} + MS_P)^2}{MS^2_{\alpha \times B}/df_{\alpha \times B} + MS^2_P/df_P}. \quad (50)$$

What Constitutes a Random-Effect?

In most behavioral science research, subjects have traditionally been treated as random effects in the ANOVA. The rationale is simple: For each investigation, researchers select a small sample from the population of interest, in order to generalize the results beyond this sample. Except in the case of summative evaluations, it is rare that anyone would be interested in the data as being presented as pertinent to only to the particular individuals studied (Santa et al., 1979). The logic that compels one to treat subjects as a random-effect, however, should also lead to treating other nonessential stimulus variation as random-effects (Clark, 1973; Coleman, 1964). That is, in many situations, educational researchers should seek

generality beyond the stimuli, settings, treatments, materials, as well as the individuals involved in a particular study.

As elaborated previously, random sampling from the population of possible conditions (i.e., settings, materials, etc.) is a condition that constitutes a random-effect. Yet for most educational researchers the question remains, how often can this be done? Clark (1976) comments that a researcher is rarely able to make an *a priori* specification of an independent variable such that each level has an equal probability of being selected. This is especially true when repeated measures are involved. If students are supposed to be measured at various times in the school year, when should these measures be taken? Obviously, the researcher wants to generalize beyond those particular points in time. Yet, time has an infinite population, and thus random sampling is not only difficult but also impractical. Keppel (1982, p. 520) states that in such cases, researchers often select levels which reflect equally spaced intervals, "inflection points," or other points of interest which, based on the researchers' experience, are assumed to be representative of the underlying continuum of the independent variable. Thus, a factor frequently is neither "fixed" nor randomly selected; however, Myers (1979, pp. 193-194) suggests that a factor should be classified as random if another investigator *could* have selected different levels of the variable. Many researchers agree that this "quasi-random" process is more reflective of the "reality" of research and does not preclude the statistical use of random effects (Clark, 1973; Santa et al., 1979). For instance, subjects are almost always treated as random effects although it is common knowledge that investigators frequently obtain convenient samples of undergraduates (Clark, 1976).

Wike and Church (1976, p. 249) took exception with Clark's analysis of random-effects and considered his ideas "unsound" and at best "questionable." That is, they did not concur with the recommendation that nonessential stimulus variation *should* be treated as a random effect to achieve generality. Wike and Church do agree with Clark, and correctly so, that if the p levels analyzed equal the P possible levels (situation f_1), then fixed effects, by definition, are analyzed. Furthermore, they also agree that in situation r (i.e., levels of the factor are selected at random from a larger pool of possible levels such that $p < P$), random effects are all too often ignored or treated as fixed. They disagree, however, in what they call situation f_2, where $p < P$ but the p level where not randomly selected. Clark (1976) and his supporters present the case that in situation f_2, designs implementing random-effects should be analyzed for generality of findings. Wike and Church adhere to the technical definition of random. Thus, they infer that Clark's definition of random and fixed

hinges on whether $p < P$, while the mode of selecting the p levels is considered unimportant or "secondary." Furthermore, they question what conclusions can be drawn from a statistical analysis based on the assumption that an f_2 situation is random and to what population these results can generalized. Cohen (1976) agrees with this sentiment stating that "random means random" and that such quasi-random processes make the distribution of the random factor rectangular (i.e., platykurtic) which can severely reduce the power of F or quasi-F since random effects are not always robust to violations of the normality assumption (Scheffe', 1959). Therefore, statistical analysis of random-effects models are considered to have inherent difficulties (Dunn & Clark, 1974). Furthermore, Wike and Church make the case that experiments under situation r could be treated as fixed as long as the generalizations are limited. One problem they state is that random selection of levels may not adequately represent the P possible levels and therefore are not "good bets" to produce the treatment effects. Thus, they feel that randomization although a good control procedure is not optimal in general. Therefore, Wike and Church emphasize the role of replication (see Rosnow & Rosenthal, 1989), rather than randomization. To elaborate, in the distinction between statistical and nonstatistical generalizations (Keppel, 1982) random effects allows a statistical, deductive generalization to the population sampled. With fixed effects in situation f_2, generalizations are limited to those particular levels; however, the researcher may want to make a nonstatistical, inductive generalization that these effects may be found under similar or even rather different circumstances. To validate this inference, however, the experiment should be replicated under other conditions (i.e., different set of experimental stimuli, different setting, etc.).

It should be noted that a variable does have to be a "control" factor in order to be treated as random. That is, the analysis of random effects applies not only to nonessential features beyond which generality is sought, but also to factors of substantive interest. For example, in studying any developmental process, most investigations use some form of a repeated measures design. It is impossible, however, for the researcher in this case to collect data at all points in time such that a fixed effect analysis fully represents "development." Even the typical treatment/control fixed effects design may have a hint of "randomness." To elaborate, the control group may indeed have none of the treatment, but how much of the treatment does the experimental group receive? Is this the only level of interest? Or will the researcher generalize beyond and between these levels or to all forms of the treatments? Furthermore, in educational and psychological research,

ethical constraints rarely allow a true control group because we cannot withhold certain treatments. In this case, the experimental group represents one level of the treatment and the control group represents a lesser level. That is, the control group does not represent "no treatment," and thus, all possible levels are not available and the limited generality of a fixed effect analysis may not be of interest. This issue of generality is analogous to a threat to construct validity referred to by Cook and Campbell (1979, p. 67) as *Confounding Constructs and Levels of Constructs*. To elaborate, an experimental manipulation may involve several discrete levels of an independent variable that has an underlying continuity. Therefore, the points on the continuum selected as the levels for a given analysis may not indicate any effect when an effect may indeed materialize if other levels are used. Thus, although treating variables as random effects to allow greater generalization often includes analyses that may be overly conservative, especially when levels are selected systematically rather than randomly (Cohen, 1976: Richter & Seay, 1986), the notion of random effects to its logical extent in many cases (i.e., continuity) actually increases power. That is, using discrete blocks to sample and represent variables which have an underlying continuum increases the difficulty in obtaining adequate sampling, reduces power, and prevents a precise analysis of interactions (Bell, 1992). Thus, if the independent variable involves a qualitative factor (i.e., sex, ethnicity, etc.), then it obviously qualifies as fixed. If there is some possible continuum of levels underlying the factor which may be of interest, then a case can be made for a random effects analysis.

Thus, there is consensus on what to do in situations f_1 and r, but in situation f_2, there are two distinct approaches:

1). Treat situation f_2 as fixed and limit generalization to the p levels analyzed. Results and conclusions, as they relate to situations and setting beyond those in the study, involve nonstatistical generalizations and inductive inference. Thus, replications are necessary to determine under which levels the treatment effect exists.

2). Treat situation f_2 as random. The findings are typically more conservative but also more generalizable. It should be noted, however, that the extent to which a quasi-random process is treated as random influences the generality of findings and needs further investigation.

I personally opt for Clark's arguments and even take them further to substantive as well as control factors. His comments forcefully remind the researcher that generalization across experimental stimuli is as important as generalization across

subjects for the effective construction of theories. Furthermore, it is comforting to know that researchers continue to be responsible for defending the generality of their findings on statistical grounds, when appropriate, but always on logical bases (Smith, 1976). Thus, if a researcher is in the dilemma of an f_2 situation, what should be considered when deciding whether or not to use a random-effect in the analysis?

1). Is the factor nonessential in that it is not of primary interest, but it also could influence the results?
2). How are the levels of the factor chosen?
3). Are the levels of the factor selected the only levels of interest?
4). Is there an underlying continuum of levels that may be of interest?
5). How large is the pool of possible conditions?
6). Are there enough selected levels of the factor to adequately represent the population of possible conditions?
7). Will generalization be made beyond and between the levels chosen or to all forms of the factor.

For any discipline of inquiry to progress, valid inferences based on appropriate generalizations of findings are necessary. Thus, these and other practical questions which will be particular to a given study should be considered when designing and analyzing educational research.

References

Bell, R. W. (1992). Continuously distributed random variables in factorial designs. *Experimental Aging Research, 18*, 47-50.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin, 57*, 49-64.

Box, G. E. P. (1954). Some theories on quadratic forms applied to the study of analysis of variance problems: Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics, 25*, 290-302.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335-359.

Clark, H. H. (1976). Reply to Wike and Church. *Journal of Verbal Learning and Verbal Behavior, 15*, 257-261

Cohen, J. (1976). Random means random. *Journal of Verbal Learning and Verbal Behavior, 15*, 261-262.

Coleman, E. B. (1964). Generalizing to a language population. *Psychological Reports, 14*, 219-226.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Davenport, J. M., & Webster, J. T. (1973). A comparison of some approximate *F*-test. *Technometrics, 15*, 779-789.

Dawes, D. E. (1983). Confessions of a methodological puritan: A response to Jackson and Jacobs. *Human Communications Research, 9*, 187-191.

Dunn, O. J., & Clark, V. A. (1974). *Applied statistics for the analysis of variance and regression.* New York: Wiley.

Forster, K. I., & Dickinson, R. G. (1976). More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F_1, F_2, F', and *min* F'. *Journal of Verbal Learning and Verbal Behavior, 15*, 135-142.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the analysis of variance and covariance. *Review of Educational Research, 42*, 237-288.

Keppel, G. (1982). *Design and analysis: A researcher's handbook* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Wadsworth.

Lindquist, E. F. (1956). *Design and analysis of experiments in psychology and education* (2nd ed.). Boston: Houghton Mifflin.

Maxwell, S. E., & Bray, J. H. (1986). Robustness of the quasi *F* statistic to violations of sphericity. *Psychological Bulletin, 99*, 416-421.

Myers, J. L. (1979). *Fundamentals of experimental design* (3rd ed.). Boston: Allyn & Bacon.

Myers, R. H., & Howe, R. B. (1971). On alternative approximate *F* tests for hypotheses involving variance components. *Biometrika, 58*, 393-396.

Richter, M. L., & Seay, M. B. (1987). ANOVA designs with subjects and stimuli as random effects: Applications to prototype effects on recognition memory. *Journal of Personality and Social Psychology, 53*, 470-480.

Rogan, J. C., & Keselman, H. J. (1977). Is the ANOVA *F*-test robust to variance heterogeneity when sample sizes are equal? An investigation via a coefficient of variation. *American Educational Research Journal, 14*, 493-498.

Rosnow, R. L. &, Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

Santa, J. L., Miller, J. J., & Shaw, M. L. (1979). Using quasi *F* to prevent alpha inflation due to stimulus variation. *Psychological Bulletin, 86*, 37-46.

Scheffe', H. (1959). *The analysis of variance.* New York: Wiley.

Smith, J. E. K. (1976) The assuming-will-make-it-so fallacy. *Journal of Verbal Learning and Verbal Behavior, 15*, 262-263.

Tomarkin, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin, 99*, 90-99.

Wampold, B. E., & Drew, C. J. (1990). *Theory and application of statistics.* New York: McGraw-Hill.

Wike, E. L., & Church, J. D. (1976). Comments on Clark's "The language-as-fixed-effect fallacy." *Journal of Verbal Learning and Verbal Behavior, 15*, 249-255.

Wilcox, R. R., Charlin, V., & Thompson, K. (1986). New Monte Carlo results on the robustness of the ANOVA, *F W*, and *F** statistics. *Communications in Statistics: Simulation and Computation, 15*, 933-944.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

Notes

1. Amidst the confusion "mixed" design refers to an analysis that involves at least one fixed and at least one random factor, rather than the common use of "mixed" as a synonym for a split-plot design which has between-subject factors crossed with within-subject factors. Mixed designs can involve all between subject factors or design with multiple repeated measures.

2. Whether time of measurement is to be considered fixed or random will be discussed later, but for the moment, time is treated as fixed (i.e., the investigator is only interested in gains at these specific time periods).

3. General guidelines for deriving the expected mean squares of an effect are as follows. The expected mean square of an effect contains: a). the variance of any random effect nested within it and b). the variance of the interaction of that effect crossed with any random effect in the model. See Keppel (1982) for an excellent review of expected mean squares. See Kirk (1982) or Winer (1971) for more advanced treatments of the topic.